Genome Organization of More Than 300 Defensin-Like Genes in Arabidopsis[1][w]

Kevin A.T. Silverstein[2], Michelle A. Graham[2], Timothy D. Paape, and Kathryn A. VandenBosch*

Department of Plant Biology, University of Minnesota, St. Paul, Minnesota 55108

Defensins represent an ancient and diverse set of small, cysteine-rich, antimicrobial peptides in mammals, insects, and plants. According to published accounts, most species' genomes contain 15 to 50 defensins. Starting with a set of largely nodule-specific defensin-like sequences (DEFLs) from the model legume *Medicago truncatula*, we built motif models to search the near-complete Arabidopsis (*Arabidopsis thaliana*) genome. We identified 317 DEFLs, yet 80% were unannotated at The Arabidopsis Information Resource and had no prior evidence of expression. We demonstrate that many of these DEFL genes are clustered in the Arabidopsis genome and that individual clusters have evolved from successive rounds of gene duplication and divergent or purifying selection. Sequencing reverse transcription-PCR products from five DEFL clusters confirmed our gene predictions and verified expression. For four of the largest clusters of DEFLs, we present the first evidence of expression, most frequently in floral tissues. To determine the abundance of DEFLs in other plant families, we used our motif models to search The Institute for Genomic Research's gene indices and identified approximately 1,100 DEFLs. These expressed DEFLs were found mostly in reproductive tissues, consistent with our reverse transcription-PCR results. Sequence-based clustering of all identified DEFLs revealed separate tissue- or taxon-specific subgroups. Previously, we and others showed that more than 300 DEFL genes were expressed in *M. truncatula* nodules, organs not present in most plants. We have used this information to annotate the Arabidopsis genome and now provide evidence of a large DEFL superfamily present in expressed tissues of all sequenced plants.

Organisms are constantly confronted with potentially pathogenic microorganisms. Yet, few encounters result in disease, due to the multilayered lines of defense each organism possesses. In vertebrates, adaptive immunity has long held center stage because of its ability to recognize almost any foreign antigen. The ancient innate immune system is equally important and provides a critical line of defense in vertebrates, invertebrates, plants, and insects (Thomma et al., 2002; Beutler, 2004; Bulet et al., 2004; Finlay and Hancock, 2004).

In plants, innate immunity occurs via elaborate mechanisms (Dangl and Jones, 2001; Veronese et al., 2003). The plant cell wall serves as a barrier to microbial penetration. Antimicrobial compounds deter would-be invaders. Should penetration occur, recognition leads to the production of reactive oxygen intermediates, cell wall strengthening, activation of protein kinase pathways, and the production of signaling intermediates. Signaling events lead to localized responses such as a hypersensitive response (programmed cell death) or to the release of antimicrobial compounds. The plant is also immunized against unrelated pathogens via systemic acquired resistance (Delaney, 1997; Dong, 2001; Gozzo, 2003).

Much attention has focused on the interaction of plant resistance genes (R-genes) and pathogenic avirulence (avr) genes (Dangl and Jones, 2001). Plants have evolved R-genes to aid in the detection of specific pathogen avr gene products. Locked in an arms race, avr and R-genes are under strong pressure to evade and reestablish detection. Arabidopsis (*Arabidopsis thaliana*) has more than 150 of the nucleotide-binding-site/Leu-rich-repeat (NBS/LRR) class of R-genes (Baumgarten et al., 2003; Meyers et al., 2003).

Considerable effort has been made to elucidate the prevalence and activity of pathogenesis-related proteins such as antimicrobial peptides (AMPs; Broekaert et al., 1997; Garcia-Olmedo et al., 1998; Theis and Stahl, 2004). AMPs are widespread throughout the plant kingdom and include thionins, defensins, lipid transfer proteins, knottins, heveins, and snakins. Each of these classes of small, cationic secreted peptides has a characteristic number and linear arrangement of Cys pairs. These Cys pairs form disulfide bridges in class-specific three-dimensional folds (Broekaert et al., 1997). Members from most of these classes are active in vitro and in transgenic plants against a broad spectrum of bacterial and fungal pathogens (Broekaert et al., 1997; Berrocal-Lobo et al., 2002). AMPs frequently exhibit tissue-specific expression in epidermal and peripheral cell layers, or are nonspecifically expressed in response to wounding and pathogen attack (Broekaert et al., 1997; Garcia-Olmedo et al., 1998).

[1] This work was supported by a National Science Foundation Plant Genome Research Program award on *Medicago truncatula* genomics (DBI no. 0110206; Principal Investigator Douglas R. Cook) and by funds from the University of Minnesota College of Biological Sciences.

[2] These authors contributed equally to the paper.

* Corresponding author; e-mail kvandenb@cbs.umn.edu; fax 612–625–1738.

[w] The online version of this article contains Web-only data.
www.plantphysiol.org/cgi/doi/10.1104/pp.105.060079.

Among the AMPs, plant defensins are particularly important. They have been identified in diverse taxa with many defense roles, including antifungal (Gao et al., 2000; Park et al., 2002; Cabral et al., 2003; Lay et al., 2003a), antibacterial (Osborn et al., 1995; Segura et al., 1998; Koike et al., 2002), anti-insect (Chen et al., 2002; Lay et al., 2003b), and protease inhibitory (Osborn et al., 1995; Wijaya et al., 2000) activities. In many cases, small differences in amino acid sequence can predict the specificity of the defense role (Garcia-Olmedo et al., 1998). In contrast to many field studies of R-genes, defensins have recently been shown to confer broad-spectrum resistance to pathogens in crops (Gao et al., 2000; Kanzaki et al., 2002).

Defensins are thought to be members of small gene families. While Arabidopsis has 15 documented defensins (Thomma et al., 2002), human and mouse have less than 50 (Schutte et al., 2002). These numbers are likely gross underestimates. Spurred by the discovery of more than 300 defensin-like Cys cluster proteins (CCPs) in the legume *Medicago truncatula* (Fedorova et al., 2002; Mergaert et al., 2003; Graham et al., 2004) and more than a dozen similar unannotated open reading frames in the Arabidopsis genome (Graham et al., 2004), we set out to systematically identify more defensin-like sequences (DEFLs) in the Arabidopsis genome and in expressed sequences of higher plants.

RESULTS

Identification of DEFLs in Arabidopsis

Our search strategy used successive iterations of hidden Markov model (HMM; Durbin et al., 1998) and BLAST (Altschul et al., 1997) searches to identify small, secreted Cys-rich peptides in plants. As a starting point for the search, we used a set of HMMs generated from legume sequences identified in earlier work (Graham et al., 2004). These sequences encoded putative CCPs with distant homology to plant defensins and scorpion toxins (Graham et al., 2004), two groups known to have antimicrobial properties (Thomma et al., 2002; Graham et al., 2004; Yount and Yeaman, 2004).

We identified 317 DEFLs in Arabidopsis, including all 15 known defensins (Figs. 1 and 2; Thomma et al., 2002). Of the 317 DEFLs, 20% were annotated at The Arabidopsis Information Resource (TAIR; Huala et al., 2001), usually as expressed or predicted proteins, and 50% were present in UniProt or the literature (Vanoosthuyse et al., 2001; Apweiler et al., 2004). Further, 20% of the total had evidence of expression prior to this work via full-length cDNAs or expressed sequence tags (ESTs), and 3% were previously predicted without expression evidence at TAIR. Approximately 10% of the identified sequences are likely pseudogenes. Details for all identified sequences are provided in Supplemental Table I and Supplemental Figure 1.

Nearly all DEFL genes are composed of two exons. The first exon (approximately 65 bp) encodes the signal peptide, and the second (approximately 200 bp) encodes the mature peptide. The average intron size for expressed and predicted DEFLs, compiled separately, is 210 bp. More than 80% of expressed and predicted DEFLs have introns between 75 and 275 bp in size (size distributions in Supplemental Fig. 2). Further, the average position of the donor splice site relative to the start-ATG in predicted sequences closely matches expressed DEFLs (68 ± 2 bp versus 66 ± 1 bp, respectively).

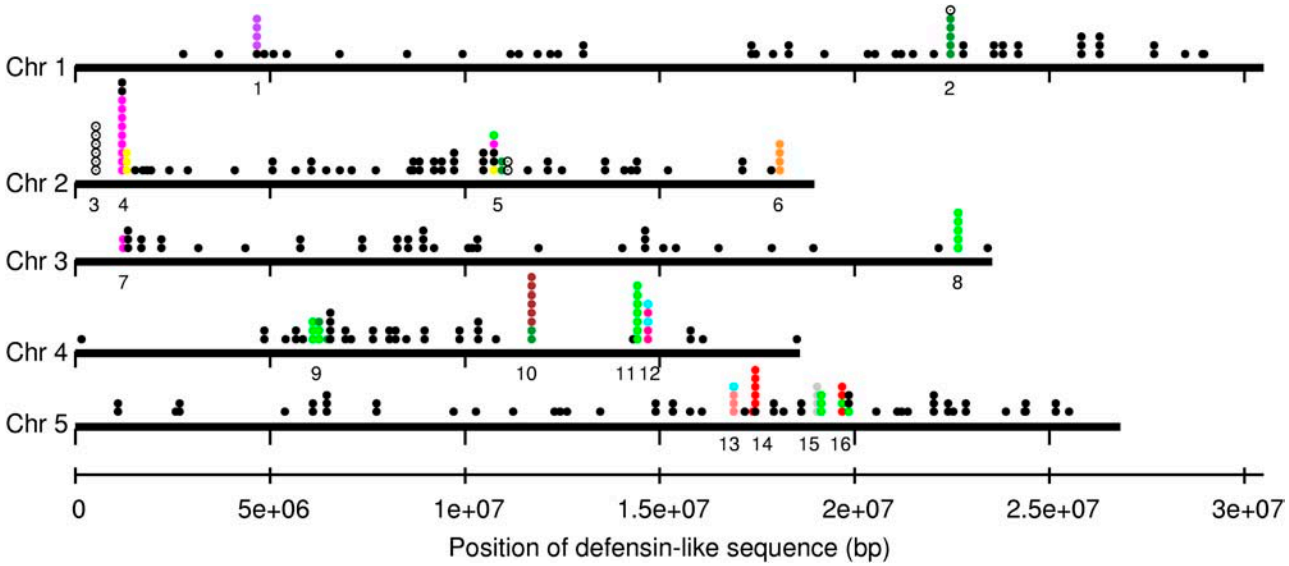

Figure 1. Clustering of related DEFLs in Arabidopsis. Each filled circle represents a single DEFL gene. DEFLs that fall within a 100,000-bp window are stacked vertically. The largest clusters are numbered for reference elsewhere. Sequences within labeled clusters are color-coded to reflect membership in sequence-related subgroups. Sequences are colored only if they match at least one other sequence of the same subgroup in a labeled cluster. Overlapping subgroups (subgroups whose members score against a neighboring subgroup's HMM with *E* value $< 10^{-4}$) are assigned a single color.

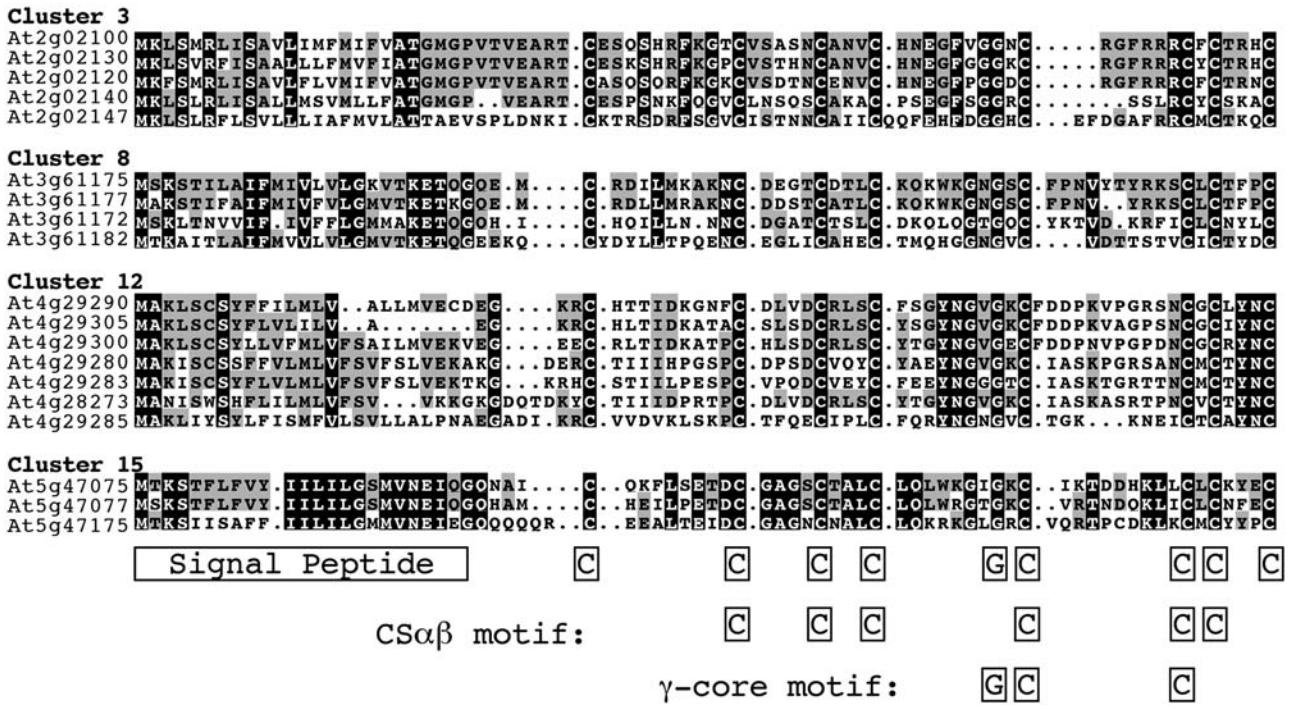

Figure 2. Representative alignment of DEFLs from four distinct clusters in Arabidopsis. Clusters correspond to those identified in Figures 1 and 3 and Table I. Not all sequences from each cluster are included. Identical residues are shaded black, while similar residues are shaded gray. A box labeled Signal Peptide designates the size and position of the signal peptide. Boxes labeled C or G designate conserved Cys or Gly residues within each group. Locus identifiers, issued by TAIR, are shown to the left of the alignment. Below the figure, key residues are shown that correspond to the CS$\alpha\beta$ motif common to defensins (Cornet et al., 1995; Lay et al., 2003b) and the γ-core motif common to all of classes Cys containing AMPs (Yount and Yeaman, 2004).

Arabidopsis DEFLs were divided into 46 subgroups, each modeled with separate HMMs. Within a subgroup, signal peptide sequence, intron position, and intron size were well conserved. For example, 75% of sequences had intron start positions within 3 bp and intron sizes within 75 bp of the subgroup average (Supplemental Figs. 3 and 4). However, the mature peptides within subgroups were highly divergent with the exception of conserved Cys. This is reflected in the predicted pIs for mature peptides within individual subgroups, where some members are quite basic and others highly acidic. Indeed, the distribution of pI values for all DEFLs is bimodal with a distinct trough around 7.0 pH units (Supplemental Fig. 5). Of these subgroups, 78% had a Cys-stabilized alpha beta (CS$\alpha\beta$) motif common to defensins (Cornet et al., 1995; Lay et al., 2003b), and 80% had a γ-core motif common to all classes Cys containing AMPs (Fig. 2; Yount and Yeaman, 2004).

Genome Organization of DEFLs in Arabidopsis

Clusters of DEFLs were explored to determine if they arose from tandem duplication and/or unequal recombination (Fig. 1; Table I). Across all clusters, nucleotide identity of repeats ranged from 48.7% to 93.9%. Repeats with low levels of nucleotide identity are likely of fairly ancient origin, while relatively recent duplications maintain a high level of nucleotide identity. The range in size of repeats (156–5,346 bp), the nucleotide identity between repeats, and the number of DEFLs per repeat (1–3) demonstrate that each cluster appears to be on its own trajectory.

In addition to duplications within clusters, we also found evidence of single or multiple gene duplications to remote sites. Sequences within closely related subgroups appear dispersed throughout the genome (Supplemental Fig. 6). Within subgroups, there is evidence for at least 80 independent segmental duplication events (Supplemental Table II), half of which involve pairs of DEFLs with >50% amino acid identity. Dotplot analysis reveals that these segmental duplications are very small (typically 500–1,500 bp) and encompass only the defensin and adjacent regulatory regions (Supplemental Table III). One stunning example is a 1,450-bp duplication on chromosomes 3 and 5. The duplicated DEFLs (At3g59930 and At5g33355) share 97% nucleotide identity throughout the signal peptide, intron, and mature protein. Once duplication to remote sites occurred, duplicated sites often underwent subsequent independent duplication or recombination events. For example, two pairs of DEFLs are duplicated on chromosomes 2 and 5 (At2g26010 and At2g26020, and At5g44430 and At5g44420). The duplicated regions, approximately 4,300 bp in length, share 79% nucleotide identity. Using the polymorphisms identified in the

Table I. *Duplication patterns observed in clusters of DEFL genes from Arabidopsis*

Cluster[a]	Cluster Size	Repeats[b]	Repeat Size	Repeat Type (Distance)[c]	Percent Identity[d]	DEFLs per Repeat
			bp			
1	5	1-1, 1-2, 1-3	436, 442, 445	D (470–688)	56.1–66.3	1
2	6	2-1 to 2-5*	408–924	D (462–1,316)	62.4–83.4	1
3	6	3-1 to 3-6	178–1,262	D (610–8,708)	48.7–76.1	1
4	14	4-1, 4-2	5,346; 5,023	D (1,535)	93.9	3
		4-3, 4-4	710, 627	D (4,285)	74.6	1
		4-5, 4-6	966, 938	D (3,103)	77.8	1
		4-7, 4-8, 4-9	479, 477, 430	D (597–769)	57.9–67.9	1
5	9	5-1, 5-2, 5-3*	1,856; 1,632; 344	T	73.1–89.3	1
6	5	6-1 to 6-5*	347–798	D (301–1,640)	49.8–71.5	1
7	5	7-1, 7-2	1,053; 935	D (675)	75.7	1
8	5	8-1*, 8-2, 8-3	382; 1,503; 1,473	T	91.5	1
		8-4, 8-5	1,315; 1,506	T	67.7	1
9	7	9-1, 9-2	1,635; 1,894	D (1,404)	75.9	1
		9-3, 9-4	1,718; 1,711	D (924)	76.9	1
10	8	10-1 to 10-6	467–780	D (598–2,374)	59.3–85.3	1
		10-7, 10-8	847, 904	D (3,673)	62.5	1
11	7	11-1, 11-2, 11-3*	721, 744, 546	D (1,403; 851)	67.5–84.5	1
		11-4, 11-5, 11-6*	2,048; 2,532; 1,897	T	69.2–89.7	1
12	6	12-1, 12-2	647, 687	D (1,157)	66	1
13	4	13-1, 13-2	843, 899	D (1,108)	77.4	1
14	6	14-1 to 14-5	1,085–3,130	D (663–2,014)	81.6–95.0	1
15	6	15-1, 15-2	7,379; 4,648	T	81	2
		15-3, 15-4	939; 1,362	T	87.9	1
16	7	16-1, 16-2, 16-3	766, 909, 618	D (8,938–13,783)	83.2–87.7	1

[a]Cluster refers to labeled clusters of DEFL genes in Figure 1. [b]Subrepeats within a cluster on separate lines. Truncated repeats marked (*). [c]T, Tandem repeat; D, dispersed repeat. Distance between dispersed repeats in parentheses. [d]Regions containing gaps were not included.

alignment, it is clear that two genes were duplicated as a unit. However, one duplicated pair has undergone subsequent unequal recombination. Interestingly, relatively few non-local related DEFL pairs overlap known large-scale segmental duplications in Arabidopsis (Supplemental Fig. 7; Vision et al., 2000; Cannon et al., 2003).

Experimental Verification of Expression in Arabidopsis

Given the high percentage of novel genes predicted in this work, we attempted to verify the expression of representatives within the six largest clusters. Of the 12 primer combinations used, nine (75%) amplified expressed DEFLs from five different clusters, two primer pairs failed to detect expression, and one primer pair failed to amplify either expressed DEFLs or the genomic DNA control (Fig. 3). Highest expression levels were detected in flower RNA, but primer combinations 3.1, 3.2, 12.2, and 15.2 also detected expression in roots and shoots.

Sequencing of cloned reverse transcription (RT)-PCR products identified 19 different DEFLs (GenBank accession nos. AY803252–AY803270). Of these, two represented alternate transcripts of the same gene (AY803263 and AY803265). We estimated that the nine primer pairs used in cloning could have amplified 27 different genes, including one predicted pseudogene. Therefore, 63% of the possible sequences were recovered. Of the 17 unique DEFLs identified, 12 had no previous evidence of expression. While this is a limited sample size, it suggests a large percentage of the 317 predicted DEFLs will be expressed.

Cloned sequences spanning introns were used to test the accuracy of our intron predictions. With the exception of the one sequence with two splice variants from cluster 4, all predicted intron boundaries were correct. Both splice variants disagreed with our prediction.

Evolution of DEFLs in Arabidopsis

To examine the evolutionary pressures acting on DEFL genes, the rates of nonsynonymous (K_a) and synonymous (K_s) substitutions were determined between 75 gene pairs representing 16 different clusters (Fig. 4; Supplemental Table IV). The ratio of these two values (K_a/K_s) is an indication of purifying ($K_a/K_s < 1$) or diversifying selection ($K_a/K_s > 1$). In Figure 4, the signal peptide appears to be largely conserved. However, pairwise comparisons using the mature protein show evidence of both purifying and diversifying selection, depending on the cluster. For the signal peptide, 36 pairwise comparisons had K_a/K_s values ranging from 0.148 to 0.723 and were statistically significant at $P \leq 0.05$ (Fig. 4; Supplemental Table IV). Only a single pairwise comparison showed significant divergent selection in the signal peptide ($K_a/K_s = 2.19$, $P = 0.023$). For the mature peptide, 23 comparisons demonstrated significant evidence of purifying selection (K_a/K_s values ranged from 0.118–0.665,

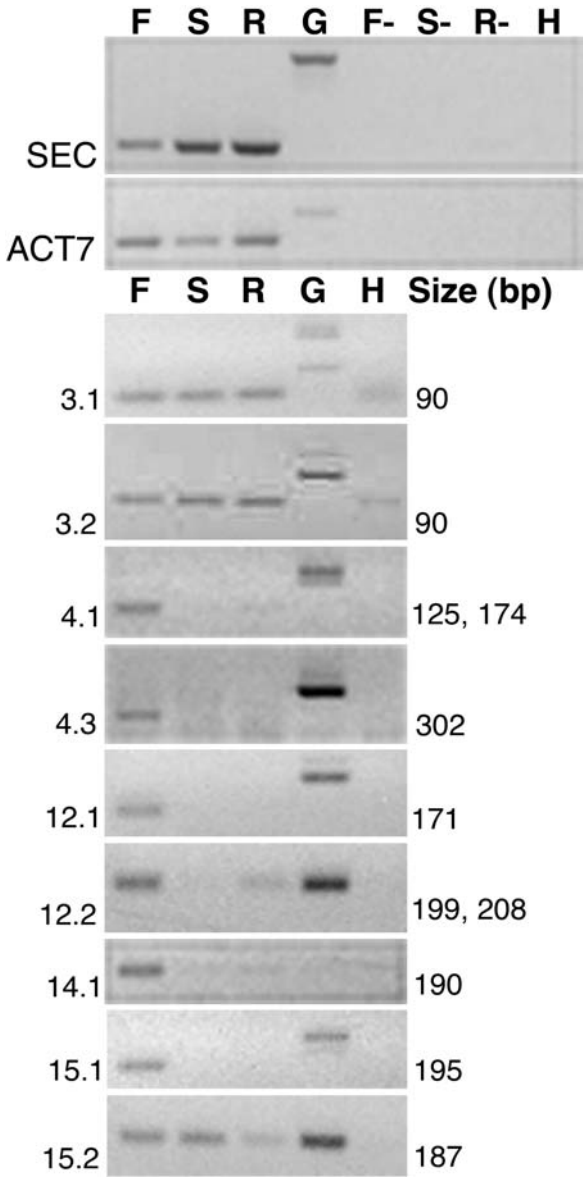

Figure 3. Expression of DEFLs in Arabidopsis flowers, shoots, and roots. Plus and minus reverse-transcriptase cDNA libraries were constructed from Arabidopsis flower (F, F−), shoot (S, S−), and root (R, R−) RNA. *SEC* and *ACT7* primers were controls for cDNA synthesis and genomic DNA contamination. Genomic DNA (G) and water (H) were positive and negative controls for PCR amplification. Primer names are located to the left of gel images and correspond to clusters in Figure 1. The sizes of the amplified cDNAs appear to the right of the image. Primer pair 14.1 was not expected to amplify genomic DNA since one of the primers spanned the boundaries of exon 1 and 2. Multiple bands within a lane reflect the ability of primers to amplify multiple DEFLs.

$P \leq 0.05$), while 18 comparisons demonstrated evidence of divergent selection (K_a/K_s values ranged from 1.406–5.239, $P \leq 0.05$).

Identification of Expressed DEFLs from Higher Plants

Outside of Arabidopsis, 1,089 unique DEFLs were identified from 62 different plant species. These sequences contributed 47 subgroups lacking an Arabidopsis counterpart. Of these, 83% had the CS$\alpha\beta$ motif, while 76% had the γ-core motif. Sequences within most of the 93 total subgroups displayed tissue-specific patterns of expression, particularly in seeds and other reproductive tissues (Table II). Note that reproductive tissues have been heavily sampled in the EST databases. They account for 41% of all plant ESTs. Thus, a subgroup of DEFLs composed almost entirely of ESTs from reproductive tissues may appear simply by chance. Despite this possibility, statistical analyses of our tissue-specific subgroups reveal that 26 of the 27 subgroups listed in Table II are indeed tissue specific ($P < 0.05$).

We also found evidence of taxon specificity among subgroups. Table II shows that 66% of all subgroups were highly specific to a single taxonomic family. In particular, many grass (Poaceae) sequences cluster into their own subgroups. Even though the Poaceae account for 53% of all unique EST sequences, the taxonomic specificity observed is statistically significant in 24 out of the 28 Poaceae-specific subgroups reported in Table II ($P < 0.05$). Expanded detail on the taxonomic and tissue distribution for ESTs in all subgroups is provided in Supplemental Table V.

DISCUSSION

Are These Genes Really Defensins?

The 317 genes described in this work have all of the hallmarks of defensin genes. Nearly all encode small putatively secreted peptides that are quite diverse with the exception of six, eight, or 10 conserved Cys. Roughly 80% have either a defensin CS$\alpha\beta$ motif (Cornet et al., 1995; Lay et al., 2003b) or a γ-core motif common to all classes of Cys-rich AMPs (Yount and Yeaman, 2004). We have shown that they have a genomic organization virtually indistinguishable from mammalian defensins (Schutte et al., 2002; Maxwell et al., 2003) and plant R-genes (Baumgarten et al., 2003; Meyers et al., 2003), which have been amplified by successive rounds of duplication and divergent selection.

Defensin-Like Genes Constitute Large Gene Families in Many Plants

Previous work suggests that defensins exist as small gene families (Schutte et al., 2002; Thomma et al., 2002). However, we have shown that the Arabidopsis genome contains 317 DEFLs. In addition, mining the collective EST data for many higher plants identified very high representation of DEFLs, particularly among the grasses. In earlier work (Fedorova et al., 2002; Mergaert et al., 2003; Graham et al., 2004), more than 300 DEFLs were identified in *M. truncatula.* The vast majority of these were expressed exclusively in the nodule, an organ not even present in Arabidopsis. Are plants truly anomalous in their number of DEFLs? It is possible. More likely, however, this highly divergent

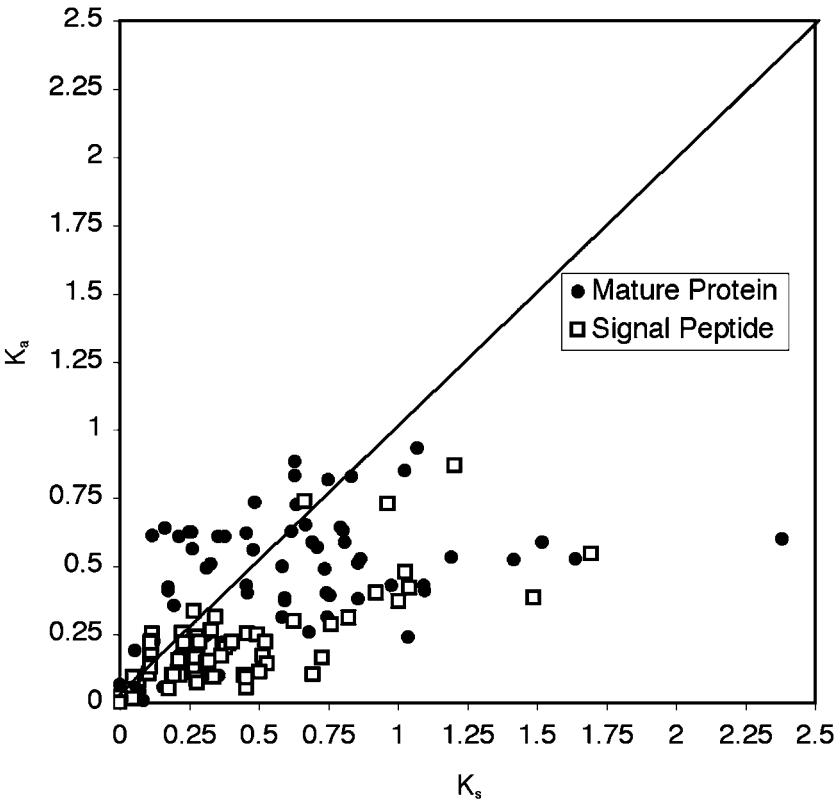

Figure 4. Plot of nonsynonymous (K_a) versus synonymous (K_s) substitution rates for Arabidopsis DEFL sequences. Nucleotide substitution ratios were determined for the coding sequences of gene pairs clearly arising from a local duplication event. The Arabidopsis clusters used in this analysis are numbered (1–16) in Figure 1. The DEFL sequences were divided into two parts: the signal peptide and the trimmed mature peptide minus conserved Cys codons. Only sequences for which a K_a and K_s value could be determined for both signal and mature peptides were included. The line represents neutral selection ($K_a/K_s = 1$). Points above this line are likely under divergent selection, while points below are conserved. Statistical analyses of these results are provided in Supplemental Table IV.

superfamily has eluded detection using current experimental and bioinformatics practices.

Current Bioinformatics Practices Have Hindered the Detection of DEFLs

The Arabidopsis genome has been nearly complete for several years. Gaps remain in centromeric and a few euchromatic regions (Hosouchi et al., 2002). Given the status of the genome sequence, it seemed surprising that such a large gene family as the DEFLs remained undiscovered. Initially, we hypothesized that current gene finding algorithms were limited in their abilities to find small genes because they were trained to recognize characteristics of known, and likely larger, genes (Zhang, 2002). Alternatively, computational data filters could have been used to remove short sequences (Scheetz et al., 2003; Wortman et al., 2003). The original Arabidopsis genome annotation suffered from the lack of accurate gene prediction software that is available today. Hence, many genes remained unpredicted, including the majority of our DEFLs. In The Institute for Genomic Research's (TIGR's) reannotation of the Arabidopsis genome, the latest suite of gene finders was utilized to capture missing gene annotations (Haas et al., 2005). A one-time minimum cutoff of 110 amino acid residues was applied to avoid adding numerous short false-positive predictions. Because of this size cutoff, the majority of our DEFLs continued to lack representation in the latest Arabidopsis genome annotation generated at TIGR (B. Haas, TIGR, personal communication).

The method we and others (Pegg and Babbitt, 1999; Vanoosthuyse et al., 2001; Schutte et al., 2002) have used is complementary to current annotation approaches. Computational size filters are often used in whole-genome annotation efforts to eliminate

Table II. *Tissue-specific and taxon-specific distribution of sequence-related subgroups from higher plants*

Tissue(s)	Subgroup Count[a]	Average Unique Sequences per Subgroup[b]	Average Redundancy per Sequence[c]
Tissue specificity			
Reproductive[d]	27	16	2.9
Seeds[d]	13	11	9.2
Nodules	2	7	5.1
Tuber skin	1	10	1.0
Roots	1	10	4.4
Taxonomic specificity			
Poaceae	28	12	7.3
Brassicaceae	19	10	0.1
Solanaceae	8	12	8.3
Fabaceae	6	14	6.3

[a]Sequence-based subgroups with ≥95% of EST sequences coming from specified taxa or tissue(s), excluding whole plant and unspecified tissues. [b]Average number of unique sequences found in TIGR's plant GIs, Uniref100, and the Arabidopsis and *M. truncatula* genomes. [c]Average EST redundancy from TIGR's plant GIs only. [d]These tissues are itemized in "Materials and Methods."

short, false-positive predictions. However, they also eliminate small genes present in the genome. Our approach offers a means to distinguish spurious predictions from families of real genes. Common features shared among predicted genes offer clues of biological importance. We and others (Pegg and Babbitt, 1999) have observed that members of large divergent superfamilies may have poor overall sequence similarity, yet have associations of biological significance. Statistical similarity between subgroups of DEFLs is very low; yet, they share similar signal peptides, Cys arrangements, and genomic organization. In our searches, we were less rigid in requiring sequence similarity but required potential hits to have an upstream signal sequence that was not built into our motif search. The consistency of the predicted donor splice site relative to the translation start site provided further validation of our predictions. This enabled a high level of confidence even prior to our experimental verification.

DEFLs Evolve by Duplication and Selection

As mentioned previously, the DEFLs in Arabidopsis exist as single genes and clusters throughout the genome. Clearly, clusters have arisen by successive rounds of local duplication. In addition, clusters have been dispersed to remote regions of the genome by segmental duplication. Within clusters, analyses of nonsynonymous and synonymous amino acid substitution rates provide evidence for evolutionary pressures that might be acting on these genes. We found that the signal peptide is conserved, while the mature peptide may be under diversifying or purifying selection depending upon the cluster analyzed. The results are similar to what has been seen in mammalian defensins (Maxwell et al., 2003; Semple et al., 2003) and the NBS/LRR family of R-genes (Baumgarten et al., 2003; Meyers et al., 2003).

The extreme divergence between subgroups and even within local clusters has made accurate sequence alignments of DEFL genes problematic, which is a requirement for accurate phylogenetic inference. However, reliable phylogenetic studies performed on NBS/LRR genes in Arabidopsis may provide insight into the evolution of DEFL genes. Baumgarten et al. (2003) and Meyers et al. (2003) found that clusters of NBS/LRR genes have evolved by successive rounds of duplication, unequal recombination, and segmental duplication to remote regions of the genome. However, these two groups disagree on the physical scale of the segmental duplications. Baumgarten et al. (2003) assert that large-scale segmental duplications and chromosomal rearrangements are responsible for the distribution of NBS/LRR genes in the genome. By contrast, Meyers et al. (2003) found that segmental duplications have occurred on a microscale level. In our analyses of the DEFL genes, our observations are closely aligned with those of Meyers et al. (2003).

Numerous DEFLs May Be Required to Protect against Potential Pathogens

Organisms are constantly confronted with potential pathogens. Therefore, it stands to reason that each organism should possess a wide range of genes to combat threats to their growth and survival. Characterized defensin peptides have been shown to have broad-spectrum activity in vitro; however, their potency is highly dependent on ionic concentrations and synergistic interactions with other AMPs (Broekaert et al., 1997; Garcia-Olmedo et al., 1998).

Another important observation is that defensins and other AMPs are often expressed in an organ- or tissue-specific manner (Broekaert et al., 1997; Garcia-Olmedo et al., 1998). Thus, considerable redundancy in function may exist: multiple gene products expressed in distinct tissues may defend against the same or overlapping sets of pathogens. In our previous work in *M. truncatula*, the majority of DEFLs were expressed in nodules. The symbiosis between plant and rhizobium leads to suppression of typical defense responses (Mithöfer, 2002; Mitra and Long, 2004). We hypothesized that nodule-specific DEFLs protect the nutrient-rich nodule from the multitude of pathogens present in the soil. Like nodules, seeds are also nutrient rich. Large amounts of protein, polysaccharides, and lipids provide energy and raw materials for germination and development of the seedling (Wang et al., 2003). When dormant, seeds may be unable to respond to biotic threats by induction of defense response genes. Therefore, developmental control of antimicrobial peptide accumulation in seeds may be a preventive measure to avert attack on nutrient-rich resources. The observation of specific expression in seeds and other reproductive tissues for many of our DEFL genes is consistent with this hypothesis.

DEFLs May Be Involved in Non-Host Resistance

With the discovery of so many DEFLs largely specific to individual plant families, one may speculate that DEFLs could be major contributors to non-host resistance. Non-host resistance is a phenomenon in which an entire plant species is resistant to a specific pathogen (Heath, 2000; Mysore and Ryu, 2004). It is believed to be a complex phenomenon involving both preformed barriers to microbial penetration and inducible defense responses. Non-host resistance provides broad-spectrum, durable protection in the field. Defensins are ideal candidates for key players in this response. They are constitutively expressed in peripheral cell layers of nutrient-rich tissues and are inducible by microbial penetration in other tissues (Broekaert et al., 1997; Garcia-Olmedo et al., 1998). They engage in complex synergistic interactions with other AMPs to increase their potency. Moreover, individual defensins are active against a broad spectrum of microbes, and have been shown to confer resistance to microbes in transgenic crops, durable over several generations

(Gao et al., 2000; Kanzaki et al., 2002). Zimmerli et al. (2004) recently showed that several defensins in Arabidopsis were up-regulated in response to the non-host pathogens responsible for barley powdery mildew and potato late blight, but not in response to closely related host pathogens. These findings are consistent with the hypothesis that defensins and related DEFLs may be major contributors to non-host resistance.

DEFLs May Have Functions Unrelated to Defense

Not all small secreted Cys-rich plant peptides have roles in defense. Some of our DEFL genes could be involved in reproductive regulation as are members of the stig1 gene family (Goldman et al., 1994), or the Sterility-locus (S-locus) Cys-rich (SCR; Schopfer et al., 1999) and related pollen coat proteins (Watanabe et al., 2000). Beginning with the male determinant of sporophytic self-incompatibility (SSI), SP11 (a SCR protein), Vanoosthuyse et al. (2001) used iterative BLAST searches to discover 37% of the peptides we identified. SP11 adopts the same three-dimensional fold as many defensins (Chookajorn et al., 2004) and displays high levels of divergent selection and allelic diversity (Watanabe et al., 2000). Binding of SP11 from self-pollen to the stigma-specific S-locus receptor kinase (*SLK*) starts the cascade of responses that leads to rejection. *SP11* and *SLK* are genetically linked at the S-locus and coevolve together (Sato et al., 2002; Chookajorn et al., 2004).

Despite the similarities between defense and pollen recognition, it is hard to see why so many SCRs would lie outside the S-locus and why they would be expressed in so many other reproductive and somatic tissues. It would make more sense that SP11 was coopted from an ancient defensin to perform a new function (Nasrallah, 2002). While defensins are widely dispersed among eukaryotes, only a limited distribution of flowering plants use SSI, suggesting that it has only recently evolved (Hiscock and McInnis, 2003).

CONCLUSION

We set out to systematically identify DEFLs in the Arabidopsis genome and in the expressed sequences of higher plants. In Arabidopsis, we experimentally confirmed the expression of a subset of these genes. Genome analysis demonstrates that this large gene family has evolved by successive rounds of tandem and segmental duplication followed by purifying or diversifying selection. Members of the DEFL superfamily are not restricted to legumes and are far more abundant and diverse than previously appreciated. Thus, DEFLs constitute excellent candidates for crop improvement.

MATERIALS AND METHODS

Upon request, all novel materials described in this publication will be made available in a timely manner for noncommercial research purposes, subject to the requisite permission from any third-party owners of all or parts of the material. Obtaining any permissions will be the responsibility of the requestor. The cloned sequences reported in this manuscript have been deposited in the GenBank database (accession nos. AY803252–AY803270). Identified DEFL sequences from the Arabidopsis (*Arabidopsis thaliana*) genome were provided to TAIR, and Arabidopsis Genome Initiative (AGI) gene codes were assigned.

Identification of DEFLs from Arabidopsis and Other Plants

Our search strategy used successive iterations of HMM builds and searches to identify small, secreted Cys-rich peptides in plants. BLAST (Altschul et al., 1997) similarity searches were also used as a complementary approach. As a starting point for the search, we used a set of HMMs generated from legume sequences identified in earlier work (Graham et al., 2004). These sequences encoded putative CCPs with distant homology to plant defensins and scorpion toxins (Graham et al., 2004), two groups known to have antimicrobial properties (Thomma et al., 2002; Graham et al., 2004; Yount and Yeaman, 2004). HMMs representing 15 groups of CCPs were chosen from that work (groups 36, 40, 41, 645, and 31.01–31.11) because they identified homologs in Arabidopsis. These initial HMMs and subsequent HMMs were constructed only from the mature peptide, not the signal sequence. The signal peptide was left out because it appears as a separate exon in genomic sequences. Although eliminating the signal sequence from the HMMs reduced the sensitivity of the models, it provided a measure of confidence because we required genome hits to have an upstream signal peptide.

Starting with this set of legume HMMs, the entire Arabidopsis genome (Huala et al., 2001) was translated in all six frames, and scanned with each of the HMMs using hmmsearch from the HMMer package version 2.2g (Durbin et al., 1998) with default parameters. The upstream sequence of all hits with $E < 10$ was manually screened using showorf (Rice et al., 2000) for the presence of a signal peptide. Signal peptides were confirmed with SignalP version 3.0 (Bendtsen et al., 2004). Splice sites for the single expected intron were predicted using the NetPlantGene server (Hebsgaard et al., 1996). If the server failed to predict a donor or acceptor, then the missing splice sites were predicted using alignment with close homologs when available. After all hits were manually examined and false-positive hits were removed, the new sequences were added to a master list of putative Arabidopsis defensins.

To pick up neighboring homologs in sequence space, BLASTP version 2.2.1 (Altschul et al., 1997) was used to scan all new protein sequences against the translated Arabidopsis genome. The parameters -G 12 -E 2 -M BLOSUM45 -F F were used to emphasize distant homologs. Each new hit was examined as above to identify the upstream signal peptide and splice sites. Finally, acceptable hits were added to the master list.

All new protein sequences in the master list were aligned via ClustalW version 1.82 (Thompson et al., 1994), and split out into subgroups using the dendrogram generated by that software. Each of the resulting rough set of sequence subgroups was then separately realigned via ClustalW and visualized using JalView (Clamp et al., 2004). The alignments were trimmed to remove the signal peptide. The trimmed alignments were then used as input for hmmbuild and hmmcalibrate (Durbin et al., 1998) using default parameters. All sequences in the master set were scanned against each of the HMMs, and shuffled around among models until each sequence scored best against its own HMM (after realigning and rebuilding any HMMs affected by sequence exchanges). At this stage, a full iteration cycle was completed, and the new HMMs were used to rescan the translated Arabidopsis genome. Successive iterations were carried out until convergence occurred (i.e. no new hits were found). The relevant details for all identified Arabidopsis sequences in this work (e.g. precise genome location; size, position, and prediction status of the intron; expression status) are provided in Supplemental Table I. The prediction status and location of each sequence are also depicted in Supplemental Figure 1.

Following convergence within the Arabidopsis genome, the search was expanded to include the unigene sequences from all 25 plant gene indices at TIGR (Quackenbush et al., 2001), the comprehensive uniref100 collection (version 2.2) of all known protein sequences (Apweiler et al., 2004), and the emerging genomic sequence data from the *Medicago truncatula* sequencing project (September 2004, http://www.medicago.org/genome). The TIGR plant gene indices used included (total unigene counts and ESTs, respectively): Arabidopsis AGI version 11.0 (45,683, 227,670), *Capsicum annuum* CaGI version 1.0 (10,712, 22,804), *Gossypium hirsutum* CGI version 5.0 (24,350, 52,818), *Chlamydomonas reinhardtii* ChrGI version 4.0 (30,339, 152,263), *Glycine*

max and *Glycine soja* GmGI version 11.0 (67,826, 333,481), *Helianthus annuus* HaGI version 3.0 (20,520, 59,426), *Hordeum vulgare* HvGI v. 8.0 (49,190, 341,924), *Lycopersicon esculentum* LeGI version 9.0 (31,012, 155,317), *Lotus japonicus* LjGI version 3.0 (28,460, 109,618), *Lactuca sativa* LsGI version 2.0 (22,185, 68,120), *Mesembryanthemum crystallinum* McGI version 4.0 (8,455, 25,640), *M. truncatula* MtGI version 7.0 (36,976, 189,714), *Nicotiana benthamiana* NbGI version 1.0 (6,118, 18,832), *Nicotiana tabacum* NtGI version 1.0 (10,232, 9,998), *Oryza sativa* OsGI version 15.0 (88,765, 272,567), *Allium cepa* OnGI version 1.0 (11,726, 19,553), *Pinus* spp. PGI version 4.0 (31,771, 125,061), *Secale cereale* RyeGI version 3.0 (5,347, 9,119), *Sorghum bicolor* SbGI version 8.0 (39,148, 187,282), *Saccharum officinarum* SoGI version 1.0 (95,884, 255,635), *Solanum tuberosum* StGI version 9.0 (32,553, 157,197), *Triticum aestivum* TaGI version 8.0 (123,807, 542,781), *Theobroma cacao* TcaGI version 1.0 (2,539, 5,981), *Vitis vinifera* VvGI version 3.1 (23,109, 132,316), and *Zea mays* ZmGI version 14.0 (56,364, 377,188). After these datasets were translated in six frames (translation excluded for the protein dataset uniref100), they were scanned with the HMMs generated above. The same procedures as described earlier were used to verify new hits, break up alignments into new subgroups, and build new HMMs. Splice site prediction was unnecessary for the gene indices and the uniref100 sequences as they only contained exon sequences. Non-plant hits (e.g. numerous insect defensins and scorpion toxins) from uniref100 were ignored. Iterations were carried out over the 25 gene indices, the uniref100 dataset, and the Arabidopsis and Medicago genomes until no new hits were found.

In determining the tissue-specificity of identified subgroups, the following terms were identified as reproductive tissue (1,383,976 ESTs): aleurone, anther, caryopsis, coleoptile, crown, ear (maize), embryo, endosperm, fiber (cotton), flower, fruit, grain, head (wheat), inflorescence, kernel, ovary, ovule, panicle, pedicel, pericarp, pistil, pod, pollen, scutellum, seed, silk (maize), silique, sperm cell, spike, and tassel. Seed tissues (558,352 ESTs) included: aleurone, caryopsis, coleoptile, embryo, endosperm, fiber (cotton), grain, kernel, pedicel, pod, and seed. Tissues of origin for all EST libraries among the TIGR's 25 plant gene indices were determined by judicious examination of the GenBank records, and merging this extracted information with the incomplete library descriptions at TIGR. The total number of ESTs from identifiable tissues was 3,336,384. The compiled list of tissue origins is available upon request.

A tabular summary of the characteristics of the sequence-related subgroups is provided in Supplemental Table V. Additionally, alignments, fasta files, HMMs, and expression summaries are available upon request from the corresponding author. Additional nodule-specific subgroups that lack close Arabidopsis homologs are not included in the supplemental files, nor have they been included in the statistics thus far. They have been reported previously (Graham et al., 2004) and are the subject of a separate expanded genomic analysis (unpublished data).

Statistical Analyses of Taxon and Tissue Specificity for DEFL Subgroups

We used a chi-squared association test (Dunn and Clark, 2001) to determine whether sequence subgroups were statistically enriched by sequences from a specific tissue or simply represented tissues that were overly sampled in the EST databases. This method was used as long as the expected observed count was at least five for each of two data sets within the subgroup: sequences derived from the chosen tissue set and those derived from other tissues. The chi-squared association test exaggerates statistical significance when expected observed counts are low; therefore, in cases where these counts were less than five, we used the Fisher exact probability test (Siegel, 1956). This protocol was also used to assess the taxon specificity of sequence subgroups with unigene counts replacing EST counts.

Analysis of Defense Peptide Sequence Motifs

The alignments for most subgroups had clear evidence of characteristic defense motifs described in the literature. We visually scanned each subgroup alignment for two of these motifs: the CS$\alpha\beta$ motif, which is common to all proteins adopting the three-dimensional knottin defensin fold (Cornet et al., 1995); and the γ-core motif, found in all known Cys-containing AMPs (Yount and Yeaman, 2004). The CS$\alpha\beta$ motif contains the residues C…CXXXC…C…CXC, where C = Cys, X = any amino acid, and … indicates a nonconserved number of amino acids. The γ-core motif is described by GXCX{3,9}C, or its enantiomeric sequence permutations CX{3,9}CXG and CX{3,9}GXC, where G = Gly and the numbers in braces indicate a range of nonconserved amino acids. In several cases, as noted in Supplemental Table V, we extended the range of nonconserved amino acids. Subgroups were considered to have a motif if the majority of the sequences within the alignment matched it.

Duplication of DEFLs in Arabidopsis

Sequences falling within a 100,000-bp window in the Arabidopsis genome were grouped. Clusters of four or more closely spaced sequences were then assigned cluster numbers and analyzed for evidence of local duplication (Fig. 1). In a few cases, additional related sequences could be identified in the surrounding regions and the window size was increased. The maximum window size used was 335,000 bp. JDotter (Brodie et al., 2004) was used to identify the existence and approximate boundaries of repeat regions within a cluster. Using the programs GAP, BESTFIT, and PILEUP (Wisconsin Package version 10.3; Accelrys, San Diego), the size, nucleotide identity, and number of DEFLs per repeat were determined.

Degenerate Primer Design to Identify Expressed Genes

To verify defensin expression and splice site predictions, six of the clusters identified in Figure 1 were chosen for RT-PCR analysis. Clusters 4, 8, 12, 14, and 15 were chosen because they contained the highest number of DEFLs, and only two sequences from these clusters had prior evidence of expression. Cluster 3 was chosen as a positive control since expressed defensins from this cluster had already been identified. To design degenerate primers that would amplify multiple genes from within a cluster, the corresponding gene sequences were aligned and conserved regions were identified using PILEUP. At nonconserved positions, an inosine was inserted in the primer. When possible, one primer was designed from the signal peptide, while the other was designed from the mature protein to encompass the predicted intron. Designing a primer from within the signal sequence was often difficult given the small size of the first exon (typically 65 bp) and its AT-rich sequence. Due to sequence diversity within a cluster, multiple primer pairs were often designed for each cluster. The primer sequences and annealing temperatures are listed in Supplemental Table VI. Similarity between the primers and predicted genes were used to determine the total number of genes that could be amplified. A 1-bp difference was allowed in the analysis.

Two primer pairs were used as controls for experimental procedures. Primers (Supplemental Table VI) secNS2/secNS3 correspond to the secret agent gene (*SEC*, At3g04240), and primers act2f/act2r2 correspond to actin 7 (*ACT7*, At5g09810). *SEC* and *ACT7* are expressed in a variety of tissues and developmental stages (TAIR; http://www/arabidopsis.org).

Plant Materials

To obtain shoot and root material for RT-PCR, Arabidopsis ecotype Columbia seeds were surface sterilized and grown on agar plates containing Murashige and Skoog salts (2.16 g L^{-1}; Sigma, St. Louis), 1% (w/v) Suc, and 0.8% (w/v) agar. Plates were chilled for 2 d at 4°C and then placed vertically in a growth chamber programmed to provide a 16-h-day/8-h-night cycle at 21°C. The light intensity of the growth chamber was approximately 80 μE m^{-2} s^{-1}. Following 10 d of growth, shoots and roots were separated and frozen in liquid nitrogen. Floral material was obtained from plants grown on soil (LG3 and LP5; Sungrow Horticulture, Bellevue, WA) at 22°C with a constant light intensity of 80 μE m^{-2} s^{-1}. Floral material ranged from immature inflorescence to flowers 2 DPA.

RT-PCR of DEFLs

Prior to RT-PCR, total RNA was isolated from root, shoot, and flower samples using the RNeasy Plant mini kit (Qiagen, Valencia, CA). To remove contaminating genomic DNA, RNA samples were treated with DNA-free (Ambion, Austin, TX). First-strand cDNA synthesis of all three samples was performed using Transcriptor reverse transcriptase (Roche, Indianapolis), Oligo-p$(dT)_{15}$ primer (Roche), and 0.25 μg of total RNA, following the manufacturer's recommendations. Minus reverse-transcriptase libraries were made from all three samples to test for genomic DNA contamination. Following cDNA synthesis, PCR was performed using a PTC-225 DNA Engine thermocycler from MJ Research (Watertown, MA). Control primers mentioned above were used to test for genomic DNA contamination and

cDNA synthesis efficiency. Once genomic DNA contamination was ruled out, the 12 primer combinations shown in Supplemental Table VI were used to monitor the expression of the DEFLs in flowers, shoots, and roots. PCR reactions were 20 μL in volume and contained 1 × Promega PCR buffer, 2.2 mM $MgCl_2$, 200 μM each dNTP, 0.2 μM each primer, 2 μL of template cDNA, and 0.5 units of Taq DNA Polymerase (Promega, Madison, WI). PCR cycling conditions were 94°C for 2 min, 35 cycles of 94°C for 45 s, anneal for 30 s, 72°C for 45 s, followed by 72°C for 7 min. In addition to the six cDNA templates, genomic DNA was used as a positive control for PCR amplification. Following PCR, products were visualized on 1.5% agarose gel. Upon analysis of the RT-PCR results, the flower RNA sample was chosen for use as template in all cloning reactions with the nine PCR primer pairs yielding visible product.

Cloning of RT-PCR Products

PCR reactions were repeated as above; however, the PCR reaction volume was doubled to 40 μL. PCR products were purified and concentrated using Microcon YM-30 centrifugal filter devices (Millipore, Billerica, MA). PCR products were cloned using the PGEM-T Easy Vector System II, following the manufacturer's recommendations. Plasmid DNA was prepared using the Qiaprep Spin Miniprep kit (Qiagen). Eight clones from each primer pair were selected for sequencing. Double-pass sequencing was performed by the Advanced Genetic Analysis Center at the University of Minnesota. RT-PCR product sequences were analyzed using the Sequencher software (Gene Codes, Ann Arbor, MI) and then imported into the PILEUP alignment containing all gene sequences from the corresponding cluster. Sequence comparisons were made to determine which gene produced the specific RT-PCR product and to determine the size and position of the intron if present. The size and position of identified intron sequences were compared to those predicted computationally in our analysis.

Evolutionary Analyses of DEFLs in Arabidopsis

To estimate the rates of nonsynonymous (K_a) and synonymous (K_s) substitutions, the predicted coding sequences of defensin-like genes were divided into two regions: the signal sequence as determined by SignalP and the mature protein. The coding sequences of the signal peptides and the predicted mature proteins from each cluster of defensin-like genes were aligned using PILEUP. In tobacco, defensins have been identified that contain a C-terminal prodomain in addition to the signal peptide and mature defensin (Lay et al., 2003a). Therefore, the predicted mature defensin was trimmed to begin at the first conserved Cys residue. Since the Cys residues are known to be well conserved (Thomma et al., 2002), the corresponding nucleotides were also removed. Predicted pseudogenes and DEFLs bearing no similarity to other genes in the cluster were removed from the alignments. In some cases, clusters were subdivided if subgroups within a cluster could not be reliably aligned. Equivalent subdivisions were used in analysis of both the signal peptide and mature protein. The K_a and K_s values for the signal peptide and the trimmed mature protein were determined using DIVERGE (Wisconsin package; Fig. 4). Two-by-two contingency tables (Supplemental Table IV) were created to determine the statistical significance of substitution rates. The significance of the deviations from neutral selection ($K_a/K_s = 1$) was assessed for all possible pairwise comparisons within a cluster (Hughes, 1999). Only pairwise comparisons in which the K_a/K_s ratio for signal peptide and mature protein could be determined were used.

Sequence data from this article have been deposited with the EMBL/GenBank data libraries under accession numbers AY803252 to AY803270.

ACKNOWLEDGMENTS

We thank Dr. Eva Huala for providing TAIR AGI codes and examining each of our gene predictions, Brian Haas for verifying whether or not gene prediction algorithms used at TAIR failed to detect DEFLs, and Drs. Chris Town and Jeff Esch for helpful discussions. Floral RNA samples were provided by Dr. David Marks (University of Minnesota, St. Paul). Primer pairs ACT7 and SEC were provided by Dr. Lynn Hartweck (University of Minnesota, St. Paul).

Received January 25, 2005; revised March 4, 2005; accepted March 9, 2005; published June 13, 2005.

LITERATURE CITED

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Res **32:** D115–D119

Baumgarten A, Cannon S, Spangler R, May G (2003) Genome-level evolution of resistance genes in *Arabidopsis thaliana*. Genetics **165:** 309–319

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol **340:** 783–795

Berrocal-Lobo M, Segura A, Moreno M, Lopez G, Garcia-Olmedo F, Molina A (2002) Snakin-2, an antimicrobial peptide from potato whose gene is locally induced by wounding and responds to pathogen infection. Plant Physiol **128:** 951–961

Beutler B (2004) Innate immunity: an overview. Mol Immunol **40:** 845–859

Brodie R, Roper RL, Upton C (2004) JDotter: a Java interface to multiple dotplots generated by Dotter. Bioinformatics **20:** 279–281

Broekaert WF, Cammue BPA, De Bolle MFC, Thevissen K, De Samblanx GW, Osborn RW (1997) Antimicrobial peptides from plants. Crit Rev Plant Sci **16:** 297–323

Bulet P, Stocklin R, Menin L (2004) Anti-microbial peptides: from invertebrates to vertebrates. Immunol Rev **198:** 169–184

Cabral KM, Almeida MS, Valente AP, Almeida FC, Kurtenbach E (2003) Production of the active antifungal *Pisum sativum* defensin 1 (Psd1) in *Pichia pastoris*: overcoming the inefficiency of the STE13 protease. Protein Expr Purif **31:** 115–122

Cannon SB, Kozik A, Chan B, Michelmore R, Young ND (2003) DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. Genome Biol **4:** R68

Chen KC, Lin CY, Kuan CC, Sung HY, Chen CS (2002) A novel defensin encoded by a mungbean cDNA exhibits insecticidal activity against bruchid. J Agric Food Chem **50:** 7258–7263

Chookajorn T, Kachroo A, Ripoll DR, Clark AG, Nasrallah JB (2004) Specificity determinants and diversification of the Brassica self-incompatibility pollen ligand. Proc Natl Acad Sci USA **101:** 911–917

Clamp M, Cuff J, Searle S, Barton GJ (2004) The Jalview Java alignment editor. Bioinformatics **20:** 426–427

Cornet B, Bonmatin JM, Hetru C, Hoffmann JA, Ptak M, Vovelle F (1995) Refined three-dimensional solution structure of insect defensin A. Structure **3:** 435–448

Dangl JL, Jones JD (2001) Plant pathogens and integrated defence responses to infection. Nature **411:** 826–833

Delaney TP (1997) Genetic dissection of acquired resistance to disease. Plant Physiol **113:** 5–12

Dong X (2001) Genetic dissection of systemic acquired resistance. Curr Opin Plant Biol **4:** 309–314

Dunn OJ, Clark VA (2001) Basic Statistics: A Primer for the Biomedical Sciences, Ed 3. John Wiley & Sons, New York

Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK

Fedorova M, van de Mortel J, Matsumoto PA, Cho J, Town CD, VandenBosch KA, Gantt JS, Vance CP (2002) Genome-wide identification of nodule-specific transcripts in the model legume *Medicago truncatula*. Plant Physiol **130:** 519–537

Finlay BB, Hancock RE (2004) Can innate immunity be enhanced to treat microbial infections? Nat Rev Microbiol **2:** 497–504

Gao AG, Hakimi SM, Mittanck CA, Wu Y, Woerner BM, Stark DM, Shah DM, Liang J, Rommens CM (2000) Fungal pathogen protection in potato by expression of a plant defensin peptide. Nat Biotechnol **18:** 1307–1310

Garcia-Olmedo F, Molina A, Alamillo JM, Rodriguez-Palenzuela P (1998) Plant defense peptides. Biopolymers **47:** 479–491

Goldman MH, Goldberg RB, Mariani C (1994) Female sterile tobacco plants are produced by stigma-specific cell ablation. EMBO J **13:** 2976–2984

Gozzo F (2003) Systemic acquired resistance in crop protection: from nature to a chemical approach. J Agric Food Chem **51:** 4487–4503

Graham MA, Silverstein KA, Cannon SB, VandenBosch KA (2004)

Computational identification and characterization of novel genes from legumes. Plant Physiol **135:** 1179–1197

Haas BJ, Wortman JR, Ronning CM, Hannick LI, Smith RR Jr, Maiti R, Chan AP, Yu C, Farzad M, Wu D, et al (2005) Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. BMC Biology **3:** 1–55

Heath MC (2000) Nonhost resistance and nonspecific plant defences. Curr Opin Plant Biol **3:** 315–319

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S (1996) Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucleic Acids Res **24:** 3439–3452

Hiscock SJ, McInnis SM (2003) Pollen recognition and rejection during the sporophytic self-incompatibility response: Brassica and beyond. Trends Plant Sci **8:** 606–613

Hosouchi T, Kumekawa N, Tsuruoka H, Kotani H (2002) Physical map-based sizes of the centromeric regions of *Arabidopsis thaliana* chromosomes 1, 2, and 3. DNA Res **9:** 117–121

Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, et al (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res **29:** 102–105

Hughes AL (1999) Adaptive Evolution of Genes and Genomes. Oxford University Press, New York

Kanzaki H, Nirasawa S, Saitoh H, Ito M, Nishihara M, Terauchi R, Nakamura I (2002) Overexpression of the wasabi defensin gene confers enhanced resistance to blast fungus (*Magnaporthe grisea*) in transgenic rice. Theor Appl Genet **105:** 809–814

Koike M, Okamoto T, Tsuda S, Imai R (2002) A novel plant defensin-like gene of winter wheat is specifically induced during cold acclimation. Biochem Biophys Res Commun **298:** 46–53

Lay FT, Brugliera F, Anderson MA (2003a) Isolation and properties of floral defensins from ornamental tobacco and petunia. Plant Physiol **131:** 1283–1293

Lay FT, Schirra HJ, Scanlon MJ, Anderson MA, Craik DJ (2003b) The three-dimensional solution structure of NaD1, a new floral defensin from *Nicotiana alata* and its application to a homology model of the crop defense protein alfAFP. J Mol Biol **325:** 175–188

Maxwell AI, Morrison GM, Dorin JR (2003) Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. Mol Immunol **40:** 413–421

Mergaert P, Nikovics K, Kelemen Z, Maunoury N, Vaubert D, Kondorosi A, Kondorosi E (2003) A novel family in *Medicago truncatula* consisting of more than 300 nodule-specific genes coding for small, secreted polypeptides with conserved cysteine motifs. Plant Physiol **132:** 161–173

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell **15:** 809–834

Mithöfer A (2002) Suppression of plant defence in rhizobia-legume symbiosis. Trends Plant Sci **7:** 440–444

Mitra RM, Long SR (2004) Plant and bacterial symbiotic mutants define three transcriptionally distinct stages in the development of the *Medicago truncatula/Sinorhizobium meliloti* symbiosis. Plant Physiol **134:** 595–604

Mysore KS, Ryu CM (2004) Nonhost resistance: How much do we know? Trends Plant Sci **9:** 97–104

Nasrallah JB (2002) Recognition and rejection of self in plant reproduction. Science **296:** 305–308

Osborn RW, De Samblanx GW, Thevissen K, Goderis IJ, Torrekens S, Van Leuven F, Attenborough S, Rees SB, Broekaert WF (1995) Isolation and characterization of plant defensins from seeds of Asteraceae, Fabaceae, Hippocastanaceae and Saxifragaceae. FEBS Lett **368:** 257–262

Park HC, Kang YH, Chun HJ, Koo JC, Cheong YH, Kim CY, Kim MC, Chung WS, Kim JC, Yoo JH, et al (2002) Characterization of a stamen-specific cDNA encoding a novel plant defensin in Chinese cabbage. Plant Mol Biol **50:** 59–69

Pegg SC, Babbitt PC (1999) Shotgun: getting more from sequence similarity searches. Bioinformatics **15:** 729–740

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res **29:** 159–164

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16:** 276–277

Sato K, Nishio T, Kimura R, Kusaba M, Suzuki T, Hatakeyama K, Ockendon DJ, Satta Y (2002) Coevolution of the S-locus genes *SRK, SLG* and *SP11/SCR* in *Brassica oleracea* and *B. rapa*. Genetics **162:** 931–940

Scheetz TE, Trivedi N, Roberts CA, Kucaba T, Berger B, Robinson NL, Birkett CL, Gavin AJ, O'Leary B, Braun TA, et al (2003) ESTprep: preprocessing cDNA sequence reads. Bioinformatics **19:** 1318–1324

Schopfer CR, Nasrallah ME, Nasrallah JB (1999) The male determinant of self-incompatibility in Brassica. Science **286:** 1697–1700

Schutte BC, Mitros JP, Bartlett JA, Walters JD, Jia HP, Welsh MJ, Casavant TL, McCray PB Jr (2002) Discovery of five conserved beta-defensin gene clusters using a computational search strategy. Proc Natl Acad Sci USA **99:** 2129–2133

Segura A, Moreno M, Molina A, Garcia-Olmedo F (1998) Novel defensin subfamily from spinach (*Spinacia oleracea*). FEBS Lett **435:** 159–162

Semple CA, Rolfe M, Dorin JR (2003) Duplication and selection in the evolution of primate beta-defensin genes. Genome Biol **4:** R31

Siegel S (1956) Nonparametric Statistics: For the Behavioral Sciences. McGraw-Hill, NewYork

Theis T, Stahl U (2004) Antifungal proteins: targets, mechanisms and prospective applications. Cell Mol Life Sci **61:** 437–455

Thomma BP, Cammue BP, Thevissen K (2002) Plant defensins. Planta **216:** 193–202

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673–4680

Vanoosthuyse V, Miege C, Dumas C, Cock JM (2001) Two large *Arabidopsis thaliana* gene families are homologous to the Brassica gene superfamily that encodes pollen coat proteins and the male component of the self-incompatibility response. Plant Mol Biol **46:** 17–34

Veronese P, Ruiz MT, Coca MA, Hernandez-Lopez A, Lee H, Ibeas JI, Damsz B, Pardo JM, Hasegawa PM, Bressan RA, et al (2003) In defense against pathogens. Both plant sentinels and foot soldiers need to know the enemy. Plant Physiol **131:** 1580–1590

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. Science **290:** 2114–2117

Wang TL, Domoney C, Hedley CL, Casey R, Grusak MA (2003) Can we improve the nutritional quality of legume seeds? Plant Physiol **131:** 886–891

Watanabe M, Ito A, Takada Y, Ninomiya C, Kakizaki T, Takahata Y, Hatakeyama K, Hinata K, Suzuki G, Takasaki T, et al (2000) Highly divergent sequences of the pollen self-incompatibility (S) gene in class-I S haplotypes of Brassica campestris (syn. rapa) L. FEBS Lett **473:** 139–144

Wijaya R, Neumann GM, Condron R, Hughes AB, Polya GM (2000) Defense proteins from seed of *Cassia fistula* include a lipid transfer protein homologue and a protease inhibitory plant defensin. Plant Sci **159:** 243–255

Wortman JR, Haas BJ, Hannick LI, Smith RK Jr, Maiti R, Ronning CM, Chan AP, Yu C, Ayele M, Whitelaw CA, et al (2003) Annotation of the Arabidopsis genome. Plant Physiol **132:** 461–468

Yount NY, Yeaman MR (2004) Multidimensional signatures in antimicrobial peptides. Proc Natl Acad Sci USA **101:** 7363–7368

Zhang MQ (2002) Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet **3:** 698–709

Zimmerli L, Stein M, Lipka V, Schulze-Lefert P, Somerville S (2004) Host and non-host pathogens elicit different jasmonate/ethylene responses in Arabidopsis. Plant J **40:** 633–646